# 6.0 Lesson Plan

- Discuss Quizzes/Answer Questions
- Correlation
- Causation
- Regression

# 6.1 More About Correlation

Correlation coefficient $r$ measures the strength of the linear association between $X$ and $Y$ values in a scatterplot.

- $r$ lies between -1 and 1, inclusive.
- $r$ equals 1 iff all points lie on a line with positive slope.
- $r$ equals -1 iff all points lie on a line with negative slope.
- non-zero $r$ does not imply a causal relationship.

Some examples of scatterplots.

.

# 6.2 Causation

Correlations are often high when some factor affects both $X$ and $Y$.

- GPA and SAT scores are both affected by IQ.
- number of hours spent listening to Rob Zombie and GPA are both affected by lifestyle.

So it is hard to argue that correlation implies causation. GPA does not cause SAT, and Rob Zombie does not hurt GPA.

But sometimes, there might be a causal link. Hours of study is probably correlated with GPA, and it seems likely to be causal.

Ecological correlations occur when $X$ or $Y$ or both is an average or a percentage for a group. Here causation is especially difficult to show.

The original link between smoking and lung cancer was an ecological correlation (Doll, 1955). The scatterplot showed the lung cancer rate against the proportion of smokers for 11 different countries.

What kinds of counfounding factors did the tobacco lobby suggest?

A study showed that cantons in Switzerland that had high literacy rates also had high suicide rates.

Is this an ecological correlation? What might be going on?

Suppose one plotted GPA against SAT score for a set of individuals. Would this be an ecological correlation?

What does a plot of the breast cancer rate against per capita fat consumption for various countries look like, and what conclusion should one draw?

# 6.3 Regression

Regression terminology:

- The response variable is labelled $Y$. This is sometimes called the dependent variable.
- The explanatory variable is labelled $X$. This is sometimes called the independent variable, or the covariate.

In multiple regression there are more than one kind of explanatory variable. For example, one might try to predict one's grade in statistics using both high school GPA and IQ.

Regression tries to fit the "best" straight line to the data. Specifically, it fits the line that minimizes the sum of the squared deviations from each point to the line, where deviation is measured in the vertical direction.

Note: This does not measure deviation as the perpendicular distance from the point to the line.

.

The mathematical model for regression assumes that:

1. Each point $(X_i, Y_i)$ in the scatterplot satisfies:

$$Y_i = a + bX_i + \epsilon_i$$

where the $\epsilon_i$ have a normal distribution with mean zero and (usually) unknown standard deviation.

2. The errors $\epsilon_i$ have nothing to do with one another. A large error does not tend to be followed by another large error, for example.

3. The $X_i$ values are measured without error. (Thus all the error occurs in the vertical direction, and we do not need to minimize perpendicular distance to the line.)

How does one find the estimates $\hat{a}$ and $hatb$ of the coefficients in the regression equation? We need to get the values that minimize the sum of the squared vertical distance.

The sum of the squared vertical distances is

$$f(a,b) = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2.$$

So take the derivative of $f(a,b)$ with respect to $a$ and $b$, set these equal to zero, and solve. One finds that:

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}; \qquad \hat{b} = \frac{\sum_{i=1}^{n} X_i Y_i - \bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - \bar{X}^2}.$$